# SiTH: Single-view Textured Human Reconstruction with Image-Conditioned Diffusion

Supplementary Material

## Contents


**6. Benchmark Description** 1

**7. Implementation Detail** 1
- 7.1. Back-view Hallucination Module . . . . . 1
- 7.2. Mesh Reconstruction Module . . . . . 2

**8. More Experimental Results** 2
- 8.1. Image Quality Comparison . . . . . 2
- 8.2. 3D Reconstruction Plugin . . . . . 3
- 8.3. More Benchmark Evaluation . . . . . 4
- 8.4. Robustness to View Angles . . . . . 4
- 8.5. Verification of Design Choices . . . . . 5
- 8.6. Additional Results . . . . . 7

**9. Discussion** 7
- 9.1. Data-driven v.s. Optimization . . . . . 7
- 9.2. Limitations . . . . . 8


## 6. Benchmark Description

We provide detailed descriptions of the CAPE [45] and the CustomHumans [22] dataset used for benchmark evaluation in our study. The CAPE dataset includes sequences of posed humans featuring 15 subjects. For evaluation purposes, ICON [73] selected 100 frames, each consisting of RGBA images from three viewpoints and an SMPL+D (vertex displacements) ground-truth mesh. We identified several limitations in the CAPE dataset (refer to Fig. 11) Firstly, there is limited diversity in human outfits, as most subjects wear tight clothing such as t-shirts and shorts. Secondly, the images are rendered from unprocessed point clouds, leading to rendering defects. Lastly, the ground-truth meshes are of low resolution and do not fully correspond to the input images. These issues suggest that experiments conducted solely on the CAPE dataset may be biased.

To ensure an unbiased evaluation, we introduced a new benchmark using the higher-quality, publicly available 3D human dataset, CustomHumans [22]. Specifically, we selected 60 textured human scans, each featuring different outfits, for evaluation. For each scan, we rendered test images from four different viewpoints. Note that we directly rasterize the textured scans to obtain the input images, ensuring that the ground-truth mesh precisely corresponds to the images. Fig. 10 showcases samples from our benchmarks, highlighting the increased diversity of the clothing.

Figure 10. **Examples of images and ground-truth scans in CustomHumans**. Our new benchmark contains diverse and challenging human scans for evaluation.

## 7. Implementation Detail

### 7.1. Back-view Hallucination Module

We detail the implementation of our image-conditioned diffusion model described in Sec. 3.1. Our model backbone is based on the Stable Diffusion image variations [2] which leverages CLIP features for cross-attention and VAE features for concatenation in image conditioning. In both training and inference, the pretrained VAE autoencoder and the CLIP image encoder are kept frozen. We initialize the diffusion U-Net's weights using the Zero-1-to-3 [41] model and create a trainable ControlNet [77] model following the default network setups but with an adjustment to the input channels. The ControlNet inputs contain 4 channels of masks and UV images with an optional 4 channels of camera view angles. The camera view angles are essential only when generating images from arbitrary viewpoints (instead of only back-view). The ControlNet model and the diffusion U-Net's cross-attention layers are jointly trained with $512 \times 512$ resolution multi-view images, rendered from the THuman2.0 dataset [75].

For each scan in THuman2.0, we render front-back image pairs from 20 camera angles, resulting in around 10k training pairs. We also randomly change the background colors for data augmentation. For training, we utilize a batch size of 16 images and set the learning rate to $4 \times 10^{-6}$

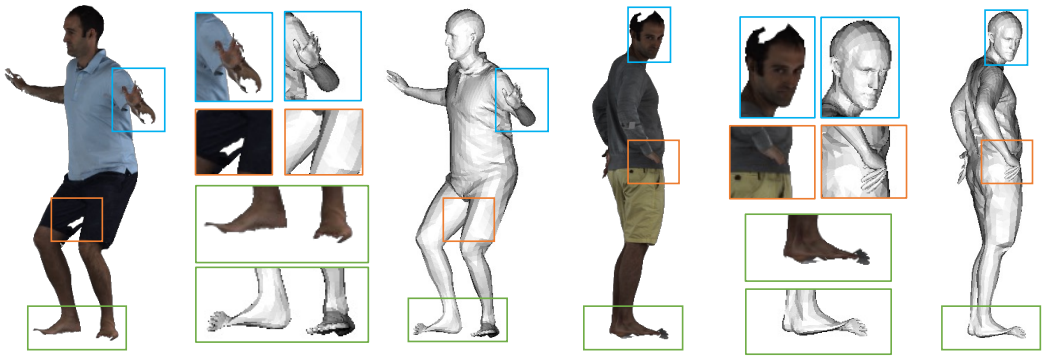

Figure 11. **Defects in the CAPE dataset**. The rendering defects from incomplete point clouds result in a notable discrepancy between the input images and the ground-truth meshes in the CAPE dataset.

incorporating a constant warmup scheduling. The ControlNet model's conditioning scale is fixed at 1.0. We employ classifier-free guidance in our training, which involves a dropout rate of 0.05 for the image-conditioning. The training takes about two days on one NVIDIA A100 GPU for 10k steps. During inference, we apply a classifier-free guidance scale of 2.5 to obtain the final output images.

### 7.2. Mesh Reconstruction Module

We follow the methodology of PIFuHD [61], using the HourGlass [50] and the fully convolutional [30] model as our image feature extractors and the normal predictor, respectively. The feature extractors yield a 32-dimensional feature map for feature querying. Our geometry MLP is designed with five layers of 512-dimensional linear layers, each followed by a leakyReLU activation function. Skip connections are applied at the third, fourth, and fifth layers. On the other hand, the texture MLP comprises four layers of 256-dimensional linear layers, with skip connections at the third and fourth layers.

We first train the normal predictor using normal images rendered from the THuman2.0 dataset. We optimize the normal predictor with an L1 reconstruction loss for 600 epochs. Subsequently, we proceed to jointly train the feature extractor and the SDF MLPs with a learning rate of 0.001 and a batch size of 2 scans. The normal predictor is jointly fine-tuned with a learning rate $1 \times 10^{-5}$. We set the hyperparameter $\lambda_n$ to 0.1. During each training iteration, we sample 40,960 query points within a thin shell surrounding the ground-truth mesh surfaces. The entire training process requires approximately five days on a single NVIDIA A100 GPU for 800 epochs on the THuman2.0 dataset. Finally, we train the other feature extractor and the RGB MLPs with a learning rate of 0.001 and a batch size of 2 scans for 200 epochs. During inference, a 3D textured mesh can be reconstructed under two minutes with an NVIDIA 3090 GPU. This includes pose estimation and mask prediction (3s), generation of a back-view image (4.5s), alignment of the body mesh and the input images (10s), and mesh reconstruction at the marching cube resolution of $512^3$ (60s).

| Method | SSIM↑ | LPIPS↓ | KID ($\times 10^{-3}$)↓ | Joints Err. (pixel)↓ |
|---|---|---|---|---|
| Pix2PixHD [69] | 0.816 | 0.141 | 86.2 | 53.1 |
| DreamPose [31] | 0.844 | 0.132 | 86.7 | 76.7 |
| Zero-1-to-3 [41] | 0.862 | 0.119 | 30.0 | 73.4 |
| ControlNet [77] +Interrogate | 0.851 | 0.202 | 39.0 | 35.7 |
| SiTH (Ours) | **0.950** | **0.063** | **3.2** | **21.5** |

Table 4. **Hallucination comparison on CustomHumans**. We compute image metrics between the generated and ground-truth back-view images. Our method achieved the best image quality and pose accuracy.

## 8. More Experimental Results

### 8.1. Image Quality Comparison

We carry out a quantitative evaluation on the images generated by **Pix2PixHD** [69], **DreamPose** [31], **Zero-1-to-3** [41], **ControlNet** [77], using ground-truth back-view images for comparison (see Tab. 4). To assess the image quality, we employ various metrics, including multi-scale Structure Similarity (**SSIM**) [70], Learned Perceptual Image Patch Similarity (**LPIPS**) [78], Kernel Inception Distance (**KID**) [6], and **2D joint errors** using a pose predictor [44]. Our method demonstrates better performance over the others in terms of similarity, quality, and pose accuracy.

In Fig. 12, we present additional results generated by these methods. DreamPose exhibits overfitting issues, failing to accurately generate back-view images with the correct appearances. Although ControlNet successfully predicts images with correct poses, it shows less accuracy in text conditioning, particularly in generating inconsistent appearances. Zero-1-to-3, shows instability in view-point conditioning, resulting in a noticeable variance in the human body poses in the generated images. In contrast, our method


Figure 12. **Qualitative comparison of back-view hallucination**. We visualize back-view images generated by the baseline methods. Note that the three different images are sampled from different random seeds. Our results are perceptually close to the ground-truth image in terms of appearances and poses. Moreover, our method also preserves generative stochasticity for handling hairstyles and clothing colors.

not only produces more faithful back-view images but also handles stochastic elements such as hairstyles and clothing colors.

### 8.2. 3D Reconstruction Plugin

We demonstrate that our hallucination module can be seamlessly integrated into existing single-view clothed human reconstruction pipelines. We implemented variants of ICON, ECON, and PIFuHD by providing them back normal from our generated back-view images (denoted as **+BH**). These are then compared to the original methods and their respective variants using the Zero-1-to-3 model as a plugin (denoted as **-123**). As shown in Fig. 13, integrating Zero-1-to-3 with these methods did not produce satisfactory clothing geometry. In contrast, our hallucination module yielded more realistic clothing wrinkles and enhanced the perceptual quality of ICON, ECON, and PIFuHD. Note that even though we provide additional images with these baselines, our pipeline still produced more detailed geometry and correct body shapes. This again verifies the importance and effectiveness of our mesh reconstruction module.

The quantitative results, presented in Table Tab. 5, further support these findings. We observed that the combination of Zero-1-to-3 with these methods did not lead to significant improvements. However, our hallucination module slightly enhanced the 3D metrics for ICON and ECON but

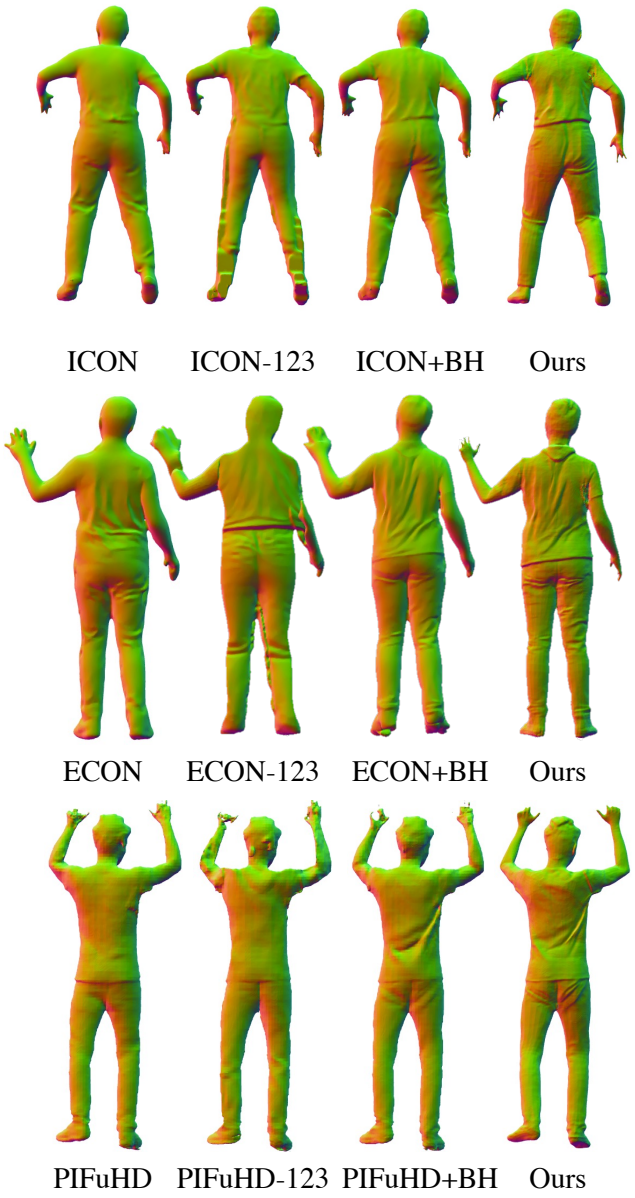


Figure 13. **Reconstruction plugin**. We replaced the back normal images typically used in existing 3D reconstruction methods with our generated back-view images. This modification enhances the perceptual qualities of these baseline methods.

had a marginally negative impact on PIFuHD. The reason can be observed from Fig. 13 where PIFuHD tends to produce smooth surfaces that result in better numeric performance. ICON and ECON benefit from our hallucinations since their original model produced artifacts and incorrect clothing details. This finding also confirms the necessity of our user studies in Sec. 4.2 since the visual quality is hard to measure by the existing metrics.

| Method | Pred-to-Scan / Scan-to-Pred (mm)↓ | NC↑ | f-Score↑ |
|---|---|---|---|
| ICON [73] | 22.562 / 27.954 | 0.791 | 30.437 |
| ICON-123 | -0.716 / -3.478 | -0.004 | +2.991 |
| ICON+BH | -1.208 / -3.728 | +0.014 | +3.831 |
| ECON [74] | 24.828 / 26.802 | 0.797 | 30.894 |
| ECON-123 | +0.646 / +1.115 | -0.024 | -1.273 |
| ECON+BH | +0.001 / +0.797 | -0.003 | -0.175 |
| PIFuHD [61] | 21.065 / 22.278 | 0.804 | 39.076 |
| PIFuHD-123 | +1.433 / +2.477 | -0.034 | -3.139 |
| PIFuHD+BH | +0.418 / -0.589 | -0.005 | +0.066 |

Table 5. **Generative plugins for 3D reconstruction**. We extend the baseline methods with Zero-1-to-3 (denoted as -123) and our hallucination module (denoted as +BH). Our method improves their perceptual qualities without affecting their overall performance. Red and blue indicate improvements and decreases respectively.

### 8.3. More Benchmark Evaluation

We present detailed descriptions of our benchmark evaluation protocol. For a fair comparison, we generated meshes from all baselines using marching cubes with a resolution of 256. To accurately compare the reconstructed meshes with ground-truth meshes, we utilize the Iterative Closest Point (ICP) algorithm [5] to register reconstructed meshes. This step is crucial for aligning the meshes with ground truth, thereby eliminating issues of scale and depth misalignment of different methods. When calculating the metrics, we sampled 100K points per mesh, and the threshold for computing the f-scores is set to 1cm. To evaluate texture reconstruction, we render front and back-view images of the generated textured meshes using aitviewer [32]. During our evaluations, we noticed that some baselines, specifically PHORHUM [4] and 2K2K [18], cannot handle non-front-facing images. Therefore, in the manuscript (Tab. 1) all the results used front-facing images. To provide a more comprehensive comparison and an evaluation aligned with real-world use cases, we include results based on images rendered from multiple view angles in Tab. 6. The CAPE and CustomHumans datasets contain images from three and four view angles respectively. Despite marginal degradation, the results indicate that our method consistently outperforms other methods in single-view 3D reconstructing.

### 8.4. Robustness to View Angles

Inspired by insights from the previous subsection, we are interested in assessing the robustness of our method against variations in image view angles. To this end, we rendered images by rotating the texture scans by $\{0, 15, 30, 45, 60, 75, 90\}$ degrees and subsequently computed their perspective 3D reconstruction metrics. This analysis is detailed in Tab. 7. We found that our pipeline

| Method | CAPE [45] | | | CustomHumans [22] | | |
|---|---|---|---|---|---|---|
| | Pred-to-Scan / Scan-to-Pred (mm)↓ | NC↑ | f-Score↑ | Pred-to-Scan / Scan-to-Pred (mm)↓ | NC↑ | f-Score↑ |
| PIFu [60] | 26.359 / 40.642 | 0.755 | 29.283 | 24.765 / 34.007 | 0.780 | 31.911 |
| PIFuHD [61] | 25.644 / 38.050 | 0.755 | 32.157 | 23.004 / 30.039 | 0.785 | **36.311** |
| FOF [15] | 21.671 / 37.246 | 0.778 | 33.971 | 21.995 / 31.076 | 0.789 | 34.403 |
| PaMIR [79] | 24.737 / 33.049 | 0.782 | 31.621 | 23.471 / 30.023 | 0.797 | 34.404 |
| ICON [73] | 27.897 / 36.907 | 0.757 | 25.898 | 25.957 / 37.857 | 0.763 | 26.857 |
| ECON [74] | 27.333 / 34.364 | 0.765 | 26.960 | 27.447 / 38.858 | 0.757 | 27.075 |
| SiTH (Ours) | **21.324 / 29.050** | **0.791** | **34.199** | **20.513 / 28.923** | **0.804** | 35.824 |

Table 6. **Single-view human reconstruction from multiple viewpoints**. We report Chamfer distance, normal consistency (NC), and f-score between ground truth and predicted meshes. Note that gray color denotes models trained on more commercial 3D human scans while the others are trained on with the public THuman2.0 dataset.

| Angle | Pred-to-Scan / Scan-to-Pred (mm)↓ | NC↑ | f-Score↑ | Δ CD | Δ NC | Δ f-Score |
|---|---|---|---|---|---|---|
| 0° | 16.880 / 20.314 | 0.8423 | 39.850 | - | - | - |
| 15° | 16.428 / 20.177 | 0.8428 | 39.971 | -0.452 / -0.137 | +0.0005 | +0.121 |
| 30° | 17.806 / 22.802 | 0.8305 | 37.154 | +0.926 / +2.488 | -0.0118 | -2.696 |
| 45° | 18.585 / 23.308 | 0.8243 | 35.652 | +1.705 / +2.994 | -0.0180 | -4.198 |
| 60° | 20.404 / 29.519 | 0.8052 | 33.675 | +3.524 / **+9.205** | -0.0371 | -6.175 |
| 75° | 22.111 / 33.309 | 0.7960 | 32.334 | +5.231 / +12.995 | -0.0463 | -7.516 |
| 90° | 23.752 / 38.338 | 0.7816 | 30.011 | +6.872 / +18.024 | -0.0607 | -9.839 |

Table 7. **Robustness of 3D reconstruction with respect to view angles**. We tested our pipeline using the images and textured scans that were rotated by varying view angles. Note that we use GT back-view images and only analyze the robustness of the mesh reconstruction module. The results from these tests demonstrate that our method maintains robustness within a view angle change of up to 45 degrees.

maintains robustness with viewpoint perturbations up to 45 degrees. However, a significant increase in the Chamfer distance was observed when the angle increased from 45 to 60 degrees. This difference could stem from potential failures in pose estimation or the underlying assumption that human bodies can be reconstructed from only front and back-view images, which may not hold true at wider angles. These observations provide a strong motivation for future research focused on enhancing the robustness of image reconstruction across varying view angles

### 8.5. Verification of Design Choices

**Image conditioning strategies.** We analyzed different strategies to incorporate image-conditioning in the diffusion U-Net. Fig. 14 depicts the effects of using the CLIP image encoder and the VAE image encoder. The results show that simply relying on the **CLIP** image encoder is not sufficient to provide accurate image conditioning. The clothing appearances cannot be accurately represented in the shared latent space of texts and images. On the other hand, the **VAE** encoder alone might also lose semantic information, such as male and female, for back-view hallucination. The hairstyles in the back are not consistent with the front-view image. Finally, the combination of both image features (**CLIP+VAE**) complements missing information of each image feature, therefore achieving more plausible results for back-view hallucination.

**ControlNet inputs.** We conducted controlled experiments to validate the efficacy of using SMPL-X UV maps and silhouette masks as conditioning inputs for our diffusion model. Fig. 15 illustrates the impact of employing different input images on the ControlNet models. Our results show that omitting the silhouette masks (**w/o Mask**) results in output images that lack consistent body shapes with the input images, especially in areas with garments like skirts. Conversely, while relying solely on silhouette masks (**w/o UV Map**) ensures shape consistency, the model struggles to differentiate between front and back views. This is particularly evident in the incorrect appearances on the head and face. Notably, the integration of both the silhouette and SMPL-X UV maps leads to more stable and accurate back-view hallucinations, thereby validating our approach.

**Parameters finetuning.** We conducted an analysis of the training strategy for our image-conditioned diffusion model by designing and comparing several training strategies.


Figure 14. **Analysis of different conditioning strategies**. We visualize the images generated by using different image features for conditioning. We show that combining both CLIP and VAE image features achieves more consistent and desirable results in back-view hallucination. Note that the four different images are sampled from different random seeds. Best viewed in color and zoom in.


Figure 15. **Analysis of different ControlNet conditions**. We visualize the images generated under various input conditions to the ControlNet model. We show that the integration of both UV maps and silhouette masks is crucial for generating spatially aligned back-view images. Note that the four different images are sampled from different random seeds. Best viewed in color and zoom in.

- We explored training from scratch, where both the diffusion U-Net and the ControlNet model are randomly initialized. This method is labeled as **From Scratch**.
- We initialized the model from a pretrained diffusion U-Net, but kept its parameters frozen during training. In this variant, only the ControlNet model's parameters are optimized, and it is denoted as **CtrlNet Only**.
- We developed a strategy that unfreezes all parameters in both the pretrained diffusion U-Net and the ControlNet model for training, referred to as **CtrlNet+U-Net Full**.
- Finally, we presented the training strategy used in our method, i.e., training only the cross-attention layers in the pretrained diffusion U-Net along with the ControlNet model (denoted as **CtrlNet+CrossAtt.**).

The results of these training strategies are depicted in Fig. 16, which shows that training a large diffusion model from scratch using only 500 3D scans is impractical. While leveraging large-scale pretraining can mitigate this issue, the CtrlNet Only training strategy fails to generate consistent appearances from front-view images. Alternatively, when we unfroze the parameters in the diffusion U-Net, the model showed improvement in generating images more aligned with the input conditional image. However, this approach led to a limitation where the model consistently produced identical output images, thus compromising its generative capability, particularly in varying clothing wrinkles and hairstyles. In contrast, our training strategy successfully generates perceptually consistent back-view images while preserving the model's generative capabilities. This strategy effectively handles the stochastic nature of clothing details and hairstyles for back-view hallucination.

| Method | Pred-to-Scan / Scan-to-Pred (mm)↓ | NC↑ | f-Score↑ |
|---|---|---|---|
| W/o Normal | **16.825 / 19.802** | **0.837** | **40.593** |
| W/ Normal | 18.709 / 20.451 | 0.826 | 37.029 |

Table 8. **Effectiveness of normal guidance**. We verify the effectiveness of incorporating normal guidance in our pipeline. While we observed that normal guidance marginally reduces performance in terms of 3D metrics, it significantly enhances the overall perceptual quality.

**Importance of normal guidance.** We validated the use of normal guidance in our mesh reconstruction module. In this experiment, we created a variant where normal images were replaced with RGB images during local feature querying. The 3D reconstruction results, as shown in Tab. 8, surprisingly indicate that this variant surpasses our model with normal guidance across all metrics. However, Fig. 17 il-

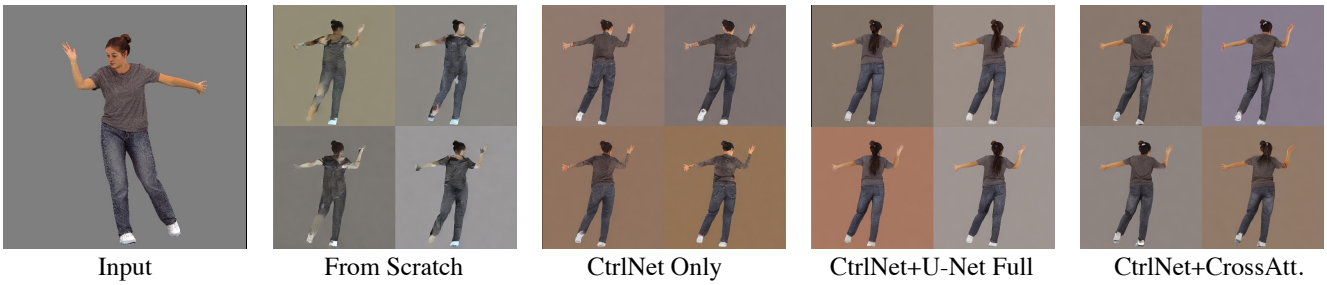


Figure 16. **Analysis of different network training strategies**. We visualize the images generated by employing different network training strategies. We show that our method produces images with consistent appearances and is able to generate diverse hairstyles and clothing details. Note that the four different images are sampled from different random seeds. Best viewed in color and zoom in.

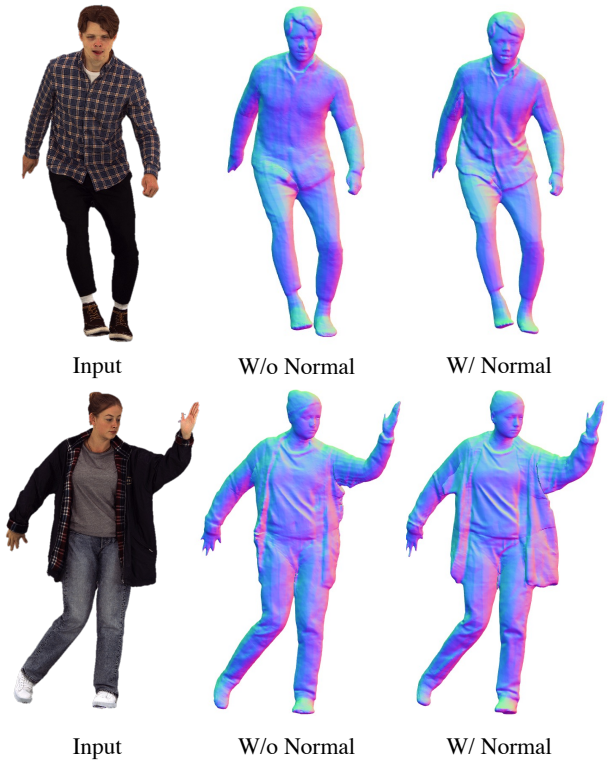


Figure 17. **Visualization of the effectiveness of normal guidance**. The use of normal guidance demonstrates its superiority in capturing geometric details of clothing and is more robust in reconstructing a challenging coat.

lustrates the tangible benefits of incorporating normal guidance. Without normal guidance, the mesh surface becomes noticeably smoother, and the model struggles to accurately reconstruct challenging clothing, such as coats. This observation aligns with our findings in Sec. 4.4 and Sec. 8.2, indicating that conventional 3D metrics may not fully capture perceptual quality. Hence, this trade-off highlights the importance of the normal predictor and guidance in achieving high-fidelity 3D human reconstruction.

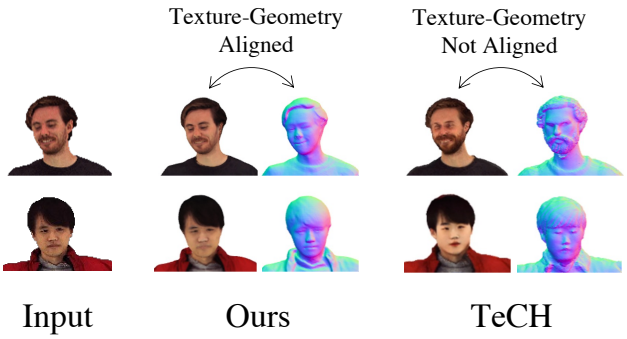


Figure 18. **Comparison with SDS optimization-based method**. Compared to the optimization-based method (TeCH), our method reconstructs consistent facial details and well-aligned mesh texture and geometry. Note that TeCH requires 6 hours to optimize both texture and geometry.

### 8.6. Additional Results

We present more qualitative results in Fig. 20, Fig. 21, Fig. 22, and Fig. 23, demonstrating our method's robustness in handling unseen images sourced from the Internet.

## 9. Discussion

### 9.1. Data-driven v.s. Optimization

Numerous concurrent works, such as TECH [24] and Human-SGD [3], propose creating 3D textured humans from single images using optimization-based approaches. These methods primarily build upon pretrained diffusion models and a Score Distillation Sampling loss, with several adaptations. In our discussion, we highlight the unique aspects of our method in comparison. Our method uniquely integrates a diffusion model into the existing data-driven 3D reconstruction workflow. This integration allows us to efficiently exploit 3D supervision to learn a generalized model for single-view reconstruction, thus avoiding the need for costly and time-consuming per-subject optimization. Consequently, our pipeline can generate high-quality textured

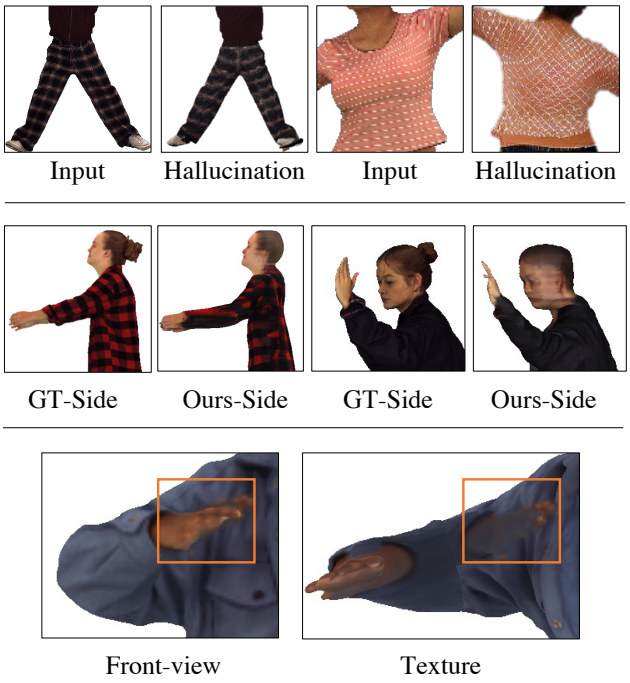


Figure 19. **Limitations**. *Top*: The hallucination model struggles with complex textures like stripes or plaid. *Middle*: Side-view appearances are not accurately recovered by mesh reconstruction. *Bottom*: The mesh reconstruction model is unable to effectively handle self-occluded regions.

meshes in under two minutes. Moreover, we observed that the existing optimization-based methods failed to generate 3D meshes having consistent and aligned texture and geometry (Fig. 18). This is due to their requirements of optimizing texture and geometry with separate optimization processes. Instead, our results are more similar to the input images and retain the consistency of both texture and geometry. Lastly, our two-stage pipeline empowers the 3D human creation process with controllability. As demonstrated in Sec. 8.1, our hallucination model handles generative stochasticity and is able to create various plausible back-view images. This feature provides users with the flexibility to choose back-view appearances based on their preferences, instead of solely relying on a random optimization process. However, as previously discussed, our method does have certain limitations. We believe that the further cross-pollination of both methods offers a promising path for future developments in generative 3D human creation.

## 9.2. Limitations

**Complex clothing textures.** We observed a challenge with the image-conditioned diffusion model in accurately generating complex clothing textures, such as stripes or plaid (Fig. 19 *Top*). This limitation stems from the image feature resolution using a pretrained VAE image encoder for feature extraction and reconstruction. The model generates output images at a resolution of $512 \times 512$, yet the diffusion U-Net is limited to processing features of only $64 \times 64$. Consequently, finer texture details may be lost in the diffusion process. This issue motivates the need for future development of pixel-perfect image-conditioning approaches, which could more accurately capture details in high-resolution images.

**Side-view appearances.** Our method follows the established practice in single-view human reconstruction, using a "sandwich-like" approach that relies on front and back information. This technique reduces the need for extensive multi-view images for 3D reconstruction. However, as shown in Fig. 19 *Middle*, a limitation of this method is the loss of detail in side views. A promising direction for future enhancements would be integrating our pipeline with optimization-based methods for a more detailed 3D human creation. Our pipeline currently provides a robust initialization by providing 3D human models with geometric and appearance details. By leveraging this initialization, the lengthy optimization process could be accelerated, making it more effective for creating detailed 3D humans.

**Self-occlusion.** Our mesh reconstruction module struggles to reconstruct appearance details in self-occluded regions, as illustrated in Fig. 19 *Bottom*. This challenge arises because essential information in these areas is not captured by either front or back-view images, and thus the mesh reconstruction module fails to infer these details. One potential solution is using an optimization process for refinement, as previously suggested. Another promising direction for future work could be developing a hallucination model capable of generating multi-view images with accurate 3D consistency, which would help reduce the self-occluded regions in mesh reconstruction.

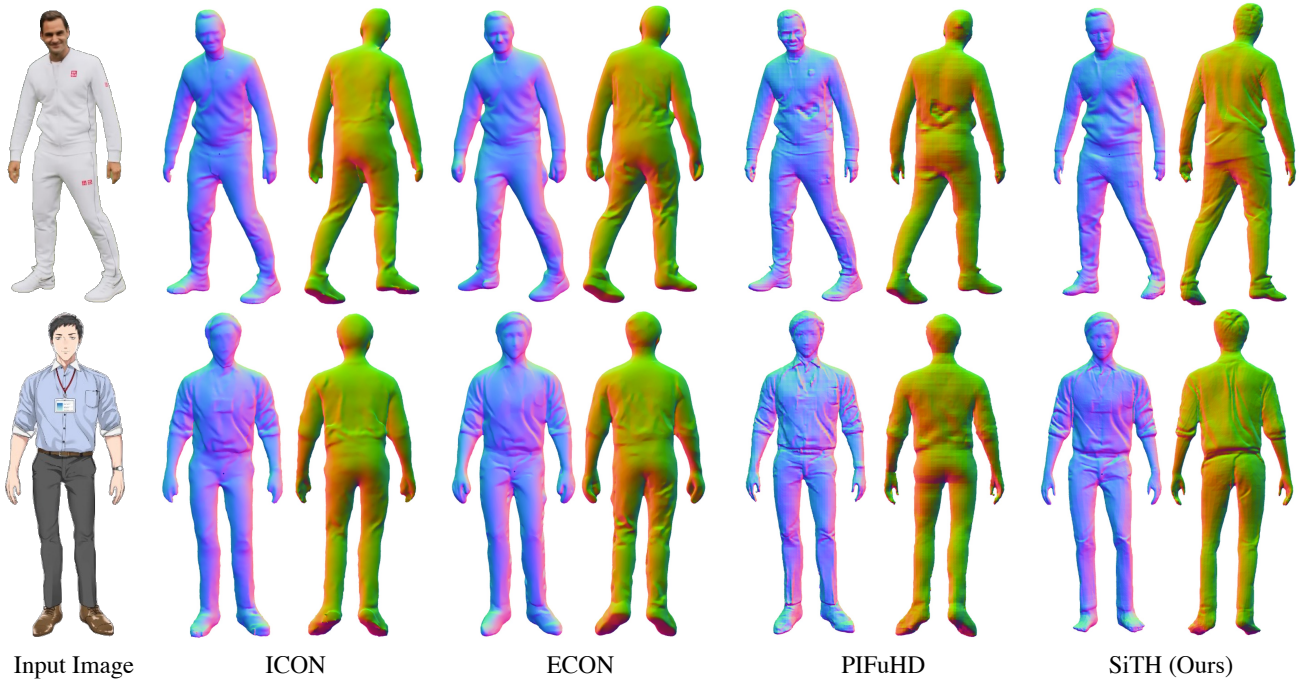


Figure 20. **Qualitative comparison of mesh geometry with Internet images**. Our method generates realistic clothing wrinkles in the back regions. Best viewed in color and zoom in.


Figure 21. **Qualitative comparison of mesh texture with Internet images**. Our method generates realistic texture in and back regions. Best viewed in color and zoom in.

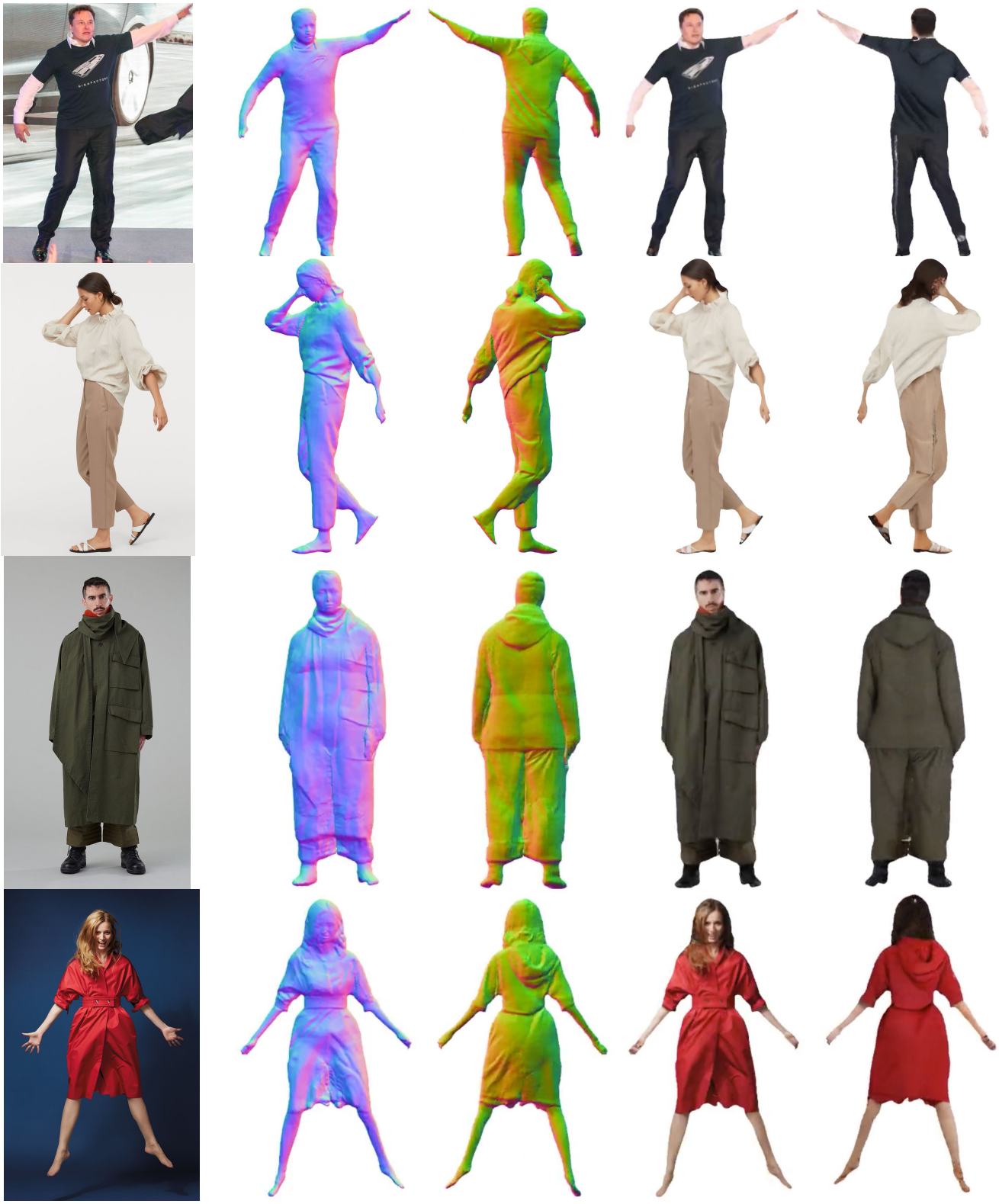

Figure 22. **Examples of reconstruction from Internet images**. Our method generates realistic clothing wrinkles in the back regions. Best viewed in color and zoom in.


Figure 23. **Examples of reconstruction from Internet images**. Our method generates realistic clothing wrinkles in the back regions. Best viewed in color and zoom in.